W205 Final Project Proposal
Team: Jonathan Landesman and Cory Kind

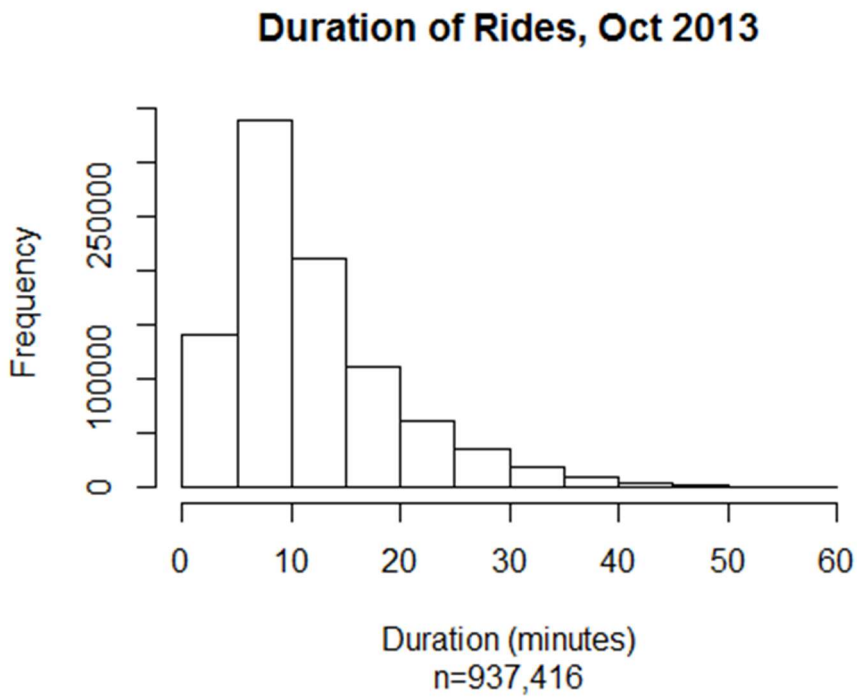
Primary Research Question
How does age affect cardiovascular ability (as demonstrated by average biking speed) among users of New York City's Citi Bike program?

We believe that average biking speed is primarily a function of distance, age, and external factors such as weather and the amount of traffic. To answer understand the nature of the relationship between age and speed, we are going to aggregate data from Citi Bike NYC and calculate the average speed of bike trips between different bike stops around the city. By narrowing our analysis to bikers who depart from a given station within a close proximity in time (i.e .within an hour, within 30 minutes, etc), we can control for factors such as traffic and weather. The data also allow us to control for gender and to get an estimate of whether the bikers are commuters (i.e. hold annual passes and depart at specific hours of the morning) or are tourists/ not dedicated bikers (i.e. hold day passes). Our bias is that commuters will be more representative of the overall general population, as commuters will be subject to greater selection bias.

Data Source
Citi Bike NYC has provided open access to the service's ride records since it opened in 2013. Downloadable files include start date, time, and location; stop date, time, and location; user type (1-day, 7-day, or annual subscriber), gender, year of birth, and bike ID. In total, there are over 22 million bike rides included in this dataset.

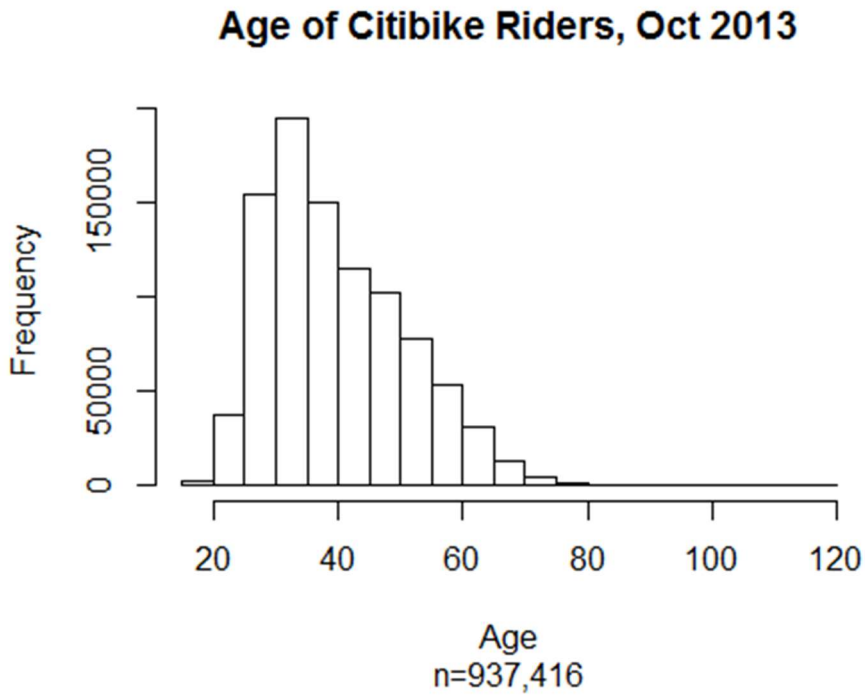
From a sample over 970,000 rows of the data from October 2013 (only lightly filtered), we can see the following results:

The median ride is under 10 minutes




```
> summary(test$duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.317   9.800  12.010  15.400  59.980
```

The median age is 38, which is older than we had initially expected:



```
summary(test$Age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.00   32.00   38.00   40.22   48.00  117.00
```

And finally, a simple univariate regression (not controlling for distance or anything else), reveals a positive, statistically significant correlation between age and trip duration. It is unlikely that older folks are riding farther distances; hence the longer duration of rides by older people are likely due to lower cardiovascular ability. This observation supports the contention that our question can be answered using the Citibike data available.

```
> summary(model1)

Call:
lm(formula = duration ~ Age, data = test)

Residuals:
    Min      1Q  Median      3Q     Max
-13.441  -5.672  -2.197   3.401  48.425

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.301653   0.031086  331.39   <2e-16 ***
Age          0.042500   0.000745   57.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.027 on 937414 degrees of freedom
Multiple R-squared:  0.00346,   Adjusted R-squared:  0.003459
F-statistic:  3255 on 1 and 937414 DF,  p-value: < 2.2e-16
```

Potential Additional Questions or Secondary Explorations

- In a research study done on the DC Capital Bikeshare system, 20% of business owners surveyed reported that adding a bikeshare station had positive impacts on sales. By combining the Citi Bike data with open Yelp data in NYC, we could try to determine if adding a bike share had a beneficial effect on the ratings of restaurants within 0.1 miles. This would obviously require combining the two datasets based on geolocation.
- If we are unable to isolate the effects of weather and traffic from the pairwise approach, we could also consider adding in that information on an hourly basis from external sources, such as NYC's Taxi data.

(Very) Rough Timeline

Week 10 & Spring Break (3/13 - 3/26)
- Estimate storage and processing needs, design architecture (changes as needed)
- Set up R in our AWS environments

Week 11 (3/27 - 4/2)
- Write ETL scripts

- Ingest and store a sample of full data set

Week 12 (4/3 - 4/9)
- Begin ingestion of full data
- Write queries to group rides by to and from location
- Explore data to understand potential cleaning issues (based on the sample, we don't anticipate this will be a significant challenge)

Week 13 (4/10 - 4/16)
- Data exploration
- Write analysis scripts to calculate differences by age among the grouped rides

Week 14 (4/17 - 4/23)
- Begin creating final report
- Finalize analyses

Week 15 (4/24 - 4/29)
- Project presentations
- Document and prepare code for submission

Potential Obstacles and Challenges

- Data do not conform to our research design: In an ideal world, we will be able to isolate rides between identical stations (station a to station b) taken within a short time span by individuals at different ages. It is possible that the flow of data will not allow such a paired design.

- Differences in data cleanliness or architecture over time: Because Citi Bike is only a few years old, there may be differences in the data structure over time that we do not anticipate.

- Selection bias: The general population self selects into Citibike customers and non-Citibike customers. Citibike customers, particularly people who use Citibike to commute, will skew towards the healthier segment of the population, and potentially exhibit lower cardiovascular decline than the general population. This bias will likely increase with age; the median Citibike customer at age 45 who uses Citibike to commute will have a health profile that is likely to be more different from the general 45 year old population than the Citibike commuting customer at age 25.